Project Description: Mean Field Asymptotics in Statistical Inference: Variational Approach, Multiple Testing, and Predictive Inference

1 Introduction

The new era of big data poses unprecedented challenges in high-dimensional statistical inference. A particular challenge that commonly exists in many inference tasks is the "dual objective" challenge, that is, the challenge of promoting efficiency and satisfying desired validity guarantees simultaneously. These two objectives compete with each other and are often difficult to be achieved simultaneously. More specifically, on the one hand, statisticians aim to design procedures that are the most statistically efficient. This is often enabled by a Bayesian component that incorporates certain prior knowledge of the scientific problem. On the other hand, statisticians hope that these procedures satisfy validity guarantees even if the assumed model and prior knowledge are wrong. Such guarantees should better be handled using frequentist approaches.

Two examples of such statistical inference tasks include false discovery rate (FDR) control [1] and predictive inference [2, 3]. In the task of FDR control, statisticians would like to design procedures that incorporate prior knowledge to make more discoveries while controlling the FDR at the specified level even when such prior knowledge is inaccurate. In the task of predictive inference, statisticians would like to give reasonably small prediction sets while maintaining the frequentist validity of coverage in the presence of model misspecification. In the past few decades, there has been a lot of progress in high dimensional statistics introducing a variety of procedures and theoretical tools towards solving these problems [1, 2, 4–18]. However, the optimal way to resolve the "dual objective" challenge in these tasks still remains open.

The first goal of this research project is to design statistical procedures that resolve the dual objective challenge in the tasks of false discovery rate control and predictive inference. The proposed procedures will be statistically near-optimal when the model is correctly specified, and satisfies the required constraints even under model misspecification. We will focus on a few stylized statistical models in the weak signal regime, in which prior knowledge and Bayesian methods should be adopted to achieve high statistical efficiency.

Moreover, in procedures with Bayesian components, variational inference methods [19], as methods for approximate Bayesian inference, are often adopted to achieve computational efficiency. Although variational inference has been widely used in many science and engineering applications [20–28], there are still many theoretical challenges in analyzing these methods, where the difficulty is largely due to the non-convexity structure of the variational objective function. Even in some well-specified statistical models with planted signals, establishing the statistical and algorithmic guarantees for many variational inference methods, such as the TAP variational inference [29], remains largely open.

The second goal of this research project is to analyze the non-convex landscape of variational inference objective functions and design efficient algorithms for optimizing these functions. We will mainly focus on the TAP variational inference, which gives consistent estimations of the Bayes posterior in models with a weak signal. Using these variational inference algorithms, we can accelerate the Bayesian inference procedures developed in our first goal.

For projects in both goals, we will build on the theoretical tools for studying high-dimensional statistical models in the asymptotic limit, which is often referred to as *mean field asymptotics* [30–45]. The mean field asymptotics theory can give sharp characterization and refined analysis of statistical procedures of interest, which is essential for the proposed research projects.

1.1 Motivations and challenges

Algorithmic and statistical properties of TAP variational inference. Variational inference methods have been widely used in various applications, including computational biology, neuroscience, computer vision, and natural language processing [20–28]. These methods include 'naive' variational Bayes, TAP

variational inference, message passing, expectation propagation [29, 46–49], etc. In recent years, there has been renewed interest in theoretical analyses of variational inference methods in high-dimensional statistical models [50–56]. However, even in some well-specified statistical models with planted signals, establishing the theoretical guarantees for many variational inference methods remains largely open.

To explain the theoretical challenge in detail, we consider a prototypical parameter estimation problem. A statistician collected a dataset $\mathscr{D}$. Assuming a parametric model $\mathscr{D} \sim \mathbb{P}_{\boldsymbol{\theta}}$ on the dataset and some prior knowledge $\boldsymbol{\theta} \sim \boldsymbol{\Pi}$, she wants to estimate the parameter $\boldsymbol{\theta} \in \mathbb{R}^d$. The Bayes optimal estimator under the square loss function is the Bayes posterior mean estimator $\boldsymbol{\theta}_{\text{Bayes}} = \mathbb{E}_{\boldsymbol{\theta}\sim\boldsymbol{\Pi},\mathscr{D}\sim\mathbb{P}_{\boldsymbol{\theta}}}[\boldsymbol{\theta}|\mathscr{D}]$. To efficiently compute the Bayes estimator $\boldsymbol{\theta}_{\text{Bayes}}$, she approximates it using $\boldsymbol{\theta}_{\text{VI}} = \arg\min_{\boldsymbol{\theta}} \mathscr{F}_{\text{VI}}(\boldsymbol{\theta})$ for some variational objective function $\mathscr{F}_{\text{VI}}$. The theoretical challenges concern how efficiently the non-convex function $\mathscr{F}_{\text{VI}}$ can be minimized, and how well $\boldsymbol{\theta}_{\text{VI}}$ approximates the Bayes estimator $\boldsymbol{\theta}_{\text{Bayes}}$.

A recent line of work analyzed the 'naive' variational Bayes method [19] for some high dimensional models with strong signals [45, 50–59]. For example, [52, 54] studied the 'naive' variational Bayes method in the stochastic block model, proving optimal error rates in the strong signal regime in which consistent recovery is possible. However, real datasets do not often possess a strong signal. Moreover, some recent results [55] show that for particular models in the weak signal regime, the 'naive' variational Bayes method does not accurately approximate the Bayes estimator. Instead, the TAP variational inference [29] could work well in these models. Nevertheless, the validity of the TAP variational inference method in many statistical models and their robustness against model misspecification remains largely open.

This research thrust aims to analyze the non-convex landscape of TAP variational inference objective functions and design efficient algorithms for optimizing these functions. We will focus on a few stylized statistical models in the weak signal regime. Our preliminary work proved some promising results for TAP variational inference in the $\mathbb{Z}_2$ synchronization problem, which is based on our new theoretical tools for analyzing the geometry of Gaussian processes. However, there are some more challenges to applying these tools to other statistical models, and we plan to resolve them in this research thrust. See Section 3 for details.

Optimal Bayesian procedures for finite-sample frequentist FDR control. Multiple testing procedures, especially the false discovery rate (FDR) control methods [1], have been widely used in biological and medical applications such as microarray experiments, functional magnetic resonance imaging, and multistage clinical trials [5, 60–63]. A fundamental question in designing these procedures is how to incorporate prior information and scientific domain knowledge to make more discoveries while controlling the FDR at the specified level even when the prior knowledge is inaccurate.

To explain the question in detail, we consider a prototypical problem of false discovery rate control. Observing a dataset $\mathscr{D}$ and assuming that $\mathscr{D}$ is generated from a parametric model $\mathbb{P}_{\boldsymbol{\theta}}$, the statistician would like to test the hypotheses that $H_{0,j} : \theta_j = 0$ for each $j = 1, 2, \ldots, d$. Her goal is to control the FDR at level α and maximize the number of discoveries. When the prior information $\boldsymbol{\theta} \sim \boldsymbol{\Pi}$ is available, the Bayesian school suggests the procedure of truncating the local FDR [5, 9], which is the posterior probability of j'th hypothesis being null $P_j \equiv \mathbb{P}_{\boldsymbol{\theta}\sim\boldsymbol{\Pi},\mathscr{D}\sim\mathbb{P}_{\boldsymbol{\theta}}}(\theta_j = 0|\mathscr{D})$, at a calibrated level calculated using the Bayes formula. In fact, when the prior $\boldsymbol{\Pi}$ is well-specified, it was shown that such a procedure is optimal [13, 60]. However, such a procedure may suffer from uncontrollable inflation of FDR when the prior information $\boldsymbol{\Pi}$ is wrong or the model $\mathbb{P}_{\boldsymbol{\theta}}$ is misspecified.

The frequentist approach of FDR control is one possible solution to make the procedure robust to certain model misspecification. One type of frequentist method calculates the p-values $(p_j)_{1\le j\le d}$ associated with some test statistics and then calibrates a rejection threshold t for these p-values using the Benjamini-Hochberg procedure [1] or its variants [4, 64, 65]. However, when the test statistics associated with the hypotheses are dependent in an unknown fashion, the Benjamini-Hochberg procedure has to be adjusted and will suffer from losing power even when the model is correctly specified. Another type of frequentist

method, including knockoff procedures [11, 12] and conditional randomization tests [12, 66], considers the model-X setting and controls the frequentist FDR with mild assumptions on the model specification. However, these methods may also lose power [67–69] compared to the oracle calibration method. It is still an open problem of finding a procedure to optimally incorporate the Bayes prior information while maintaining the frequentist FDR control in the presence of model misspecification. This is the aforementioned "dual objective" challenge in the FDR control problem.

This research thrust aims to design procedures that maximize the number of discoveries when models are correctly specified while controlling the frequentist FDR even under model misspecification. We will focus on the Bayesian linear model as a prototypical model of study. Our preliminary work applied statistical physics calculations for a few statistical procedures of interest and derived a procedure with both desired properties in the linear model with i.i.d. Gaussian design. One particularly challenging task is to design near-optimal procedures in models with anisotropic Gaussian design. See Section 4 for details.

Bayesian methods and distribution-free predictive inference. Predictive inference, with the goal to quantify the uncertainty of predictions, is an essential task in many sciences and engineering applications [3]. For example, in medical treatment recommendations, it is crucial to characterize the uncertainty associated with each prognosis. A key challenge of predictive inference is to give reasonably small prediction sets while maintaining the frequentist validity of coverage in the presence of model misspecification.

To explain the challenge in detail, we consider a prototypical problem of predictive inference. The statistician observes the dataset $\mathscr{D} = \{(\boldsymbol{x}_i, y_i)\}_{i\in[n]} \subseteq \mathscr{X} \times \mathscr{Y}$. For a new $(\boldsymbol{x}, y) \in \mathscr{X} \times \mathscr{Y}$, the statistician got the covariates $\boldsymbol{x}$, and she would like to construct a prediction set $\widehat{C}(\boldsymbol{x}) \subseteq \mathscr{Y}$ such that the chance of y is in the set $\widehat{C}(\boldsymbol{x})$ is higher than a specified threshold $1-\alpha$. She would also hope that the prediction set is as small as possible, so that the set is more informative to the actual response y. If the statistician is a Bayesian, the task would be relatively simple. The statistician could assume $(\boldsymbol{x}, y), (\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n) \sim_{iid} \mathbb{P}_{\boldsymbol{\theta}}$ and assume a prior $\boldsymbol{\theta} \sim \boldsymbol{\Pi}$, and then calculate the posterior density $p(y|\mathscr{D}, \boldsymbol{x})$. If $\mathscr{Y} = \mathbb{R}$ and the set size is measured using the Lebesgue measure, Bayesian decision theory [70] suggests that the Bayes optimal solution is $\widehat{C}_{\text{Bayes}}(\boldsymbol{x}) = \{y \in \mathbb{R} : p(y|\mathscr{D}, \boldsymbol{x}) \geq t\}$, where t is such that $\mathbb{P}(y \in \widehat{C}_{\text{Bayes}}(\boldsymbol{x})|\mathscr{D}, \boldsymbol{x}) = 1-\alpha$. Although this gives a simple solution, the Bayes prediction set may not have a valid coverage guarantee when the prior information $\boldsymbol{\Pi}$ is wrong or the model $\mathbb{P}_{\boldsymbol{\theta}}$ is misspecified.

As a predictive inference method, conformal prediction [2] attempts to resolve such a problem by giving prediction bands with marginal coverage guarantees under general exchangeability assumptions. More specifically, suppose that the statistician used a base prediction method to learn a sequence of nested prediction sets from the training dataset. Then conformal prediction methods can wrap around the base prediction method and select one set with calibrated coverage probability using the calibration dataset. The conformalized prediction set will have valid coverage guarantees under any statistical model of the i.i.d. dataset, even if the base prediction method made a wrong assumption upon the model. However, for particular tasks and datasets, it is still unclear how to choose the base prediction method so that the conformalized prediction set is the most informative when the model is correctly specified. This is the aforementioned "dual objective" challenge in the predictive inference problem.

This research thrust aims to design predictive inference methods that give reasonably small prediction sets while maintaining the frequentist validity of coverage in the presence of model misspecification. The Bayes linear model will serve as a starting point of the study, and one particularly challenging task is to understand the role of misspecified prior. We will extend the analysis to Bayesian neural networks, which were shown to perform well in many modern prediction tasks and datasets. See Section 5 for details.

1.2 Theoretical tools for analysis

All thrusts in this proposal contain tasks of calculating the exact asymptotic limit, often referred to as the "mean field asymptotics", of high dimensional statistical models in the weak signal regime. The replica method and the cavity method are the simplest heuristic methods to calculate the exact asymptotic limit, which were initially proposed for analyzing the spin glass models [71, 72] in the statistical physics literature. Due to the imminent need for big data analytics, theoretical tools for making rigorous these methods have flourished in recent years [30–45]. These theoretical tools include the approximate message passing [31, 32, 38, 41, 42], the leave-one-out approach [34, 35], the Kac-Rice formula [33, 44, 45], Gaussian comparison inequalities [36, 37, 39], and interpolation methods [30, 73–75].

2 Intellectual Merit

The proposed project has two complementary goals: 1) in the tasks of false discovery rate (FDR) control and predictive inference, design procedures that obtain statistical efficiency and satisfy desired validity guarantees simultaneously; 2) develop efficient variational inference algorithms with theoretical guarantees. Focusing on a few stylized problems and backed by extensive preliminary results, the proposed program consists of three major research thrusts.

- **Algorithmic and statistical properties of TAP variational inference.** We will analyze the non-convex landscape of TAP variational inference objective functions and design efficient algorithms for optimizing these functions. The models to be studied include spiked matrix models and Bayesian linear models, in which TAP variational inference gives consistent estimations of the Bayes posterior. We will focus on the scenario of misspecified prior and estimated prior and figure out the conditions under which the TAP objective function has a benign landscape and can be efficiently optimized.
- **Optimal Bayesian procedures for finite-sample frequentist FDR control.** In the task of FDR control, we will design procedures that maximize the number of discoveries when models are correctly specified while controlling the frequentist FDR even under model misspecification. The models to be studied include Bayesian linear and generalized linear models with anisotropic design. We will derive the exact asymptotics of the TPP and FDP tradeoff curves for the proposed procedures and compare their power with other existing procedures in the literature.
- **Bayesian methods and distribution-free predictive inference.** In the task of predictive inference, we will design procedures that give reasonably small prediction sets while maintaining the frequentist validity of coverage in the presence of model misspecification. The proposed procedures will combine Bayesian approaches with conformal prediction methods. We will further analyze Bayesian neural networks for predictive inference, which were shown to give small prediction sets in many modern prediction tasks.

This research will extend existing tools for studying the mean field asymptotics of high-dimensional statistical models. We have demonstrated this in recent work [45, 76], which extended the technique of the Kac-Rice formula and Gaussian comparison inequalities to analyze the landscape of Gaussian processes. New theoretical tools developed in this research will likely be applicable beyond the specific statistical models and problems relevant in other areas of science and engineering.

3 Component A: Algorithmic and statistical properties of TAP variational inference.

This thrust explores the algorithmic and statistical properties of variational inference methods and their robustness subject to prior misspecification. We start by analyzing the rank-1 spiked matrix model [77–79],

which is widely applied in many modern sciences and engineering applications [80]. In the spiked matrix model, we wish to estimate an n-dimensional real vector $\boldsymbol{\theta} \in \mathbb{R}^n$ having the entry-wise prior $\theta_i \sim_{iid} \Pi$. For a signal-to-noise parameter $\lambda > 0$, we observe an n by n matrix

$$\boldsymbol{Y} = (\lambda/n)\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}} + \boldsymbol{W}, \qquad \boldsymbol{W} \sim \mathrm{GOE}(n). \tag{1}$$

Here $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ is a symmetric matrix with iid entries, having off-diagonal entries $\mathcal{N}(0, 2/n)$ and diagonal entries $\mathcal{N}(0, 1/n)$. When the prior Π is symmetric, the parameter $\boldsymbol{\theta}$ is identifiable only up to $\pm$ sign, and the posterior law $p(\boldsymbol{\theta}|\boldsymbol{Y})$ has the corresponding sign symmetry $p(\boldsymbol{\theta}|\boldsymbol{Y}) = p(-\boldsymbol{\theta}|\boldsymbol{Y})$. To avoid dealing with such subtlety for symmetric priors, we will consider estimating the sign-invariant rank-one matrix $\boldsymbol{\Theta} = \boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}} \in \mathbb{R}^{n\times n}$. The Bayes posterior-mean estimate of this matrix is $\boldsymbol{\Theta}_{\mathrm{Bayes}} = \mathbb{E}[\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}} \mid \boldsymbol{Y}] \in \mathbb{R}^{n\times n}$.

The exact computation of $\boldsymbol{\Theta}_{\mathrm{Bayes}}$ is inefficient since it involves an n-dimensional integration. Therefore, statisticians often use variational inference methods to approximately compute this quantity. In the spiked matrix model, the natural approach to be considered is the TAP variational inference [29], which is conjectured to give a consistent approximation to the Bayes estimator $\boldsymbol{\Theta}_{\mathrm{Bayes}}$ [55]. The expression of the TAP free energy (the term 'free energy' is a physics terminology which could be understood here as 'objective function') of the spiked matrix model $\mathscr{F}_{\mathrm{TAP}} : \mathbb{R}^n \times \mathscr{P}(\mathbb{R}) \to \mathbb{R}$ is given as below

$$\begin{aligned} \mathscr{F}_{\mathrm{TAP}}(\boldsymbol{m}, \Pi) &= -\sum_{i=1}^{n} \mathsf{h}(m_i; Q(\boldsymbol{m}), \Pi) - \frac{\lambda}{2}\langle \boldsymbol{m}, \boldsymbol{Y}\boldsymbol{m}\rangle - \frac{n\lambda^2}{4} Q(\boldsymbol{m})^2, \\ \mathsf{h}(m; Q, \Pi) &= \inf_{\tau} \Big[\log \int_{\mathbb{R}} e^{\tau(\sigma - m) - \lambda^2 Q\sigma^2/2} \Pi(\mathrm{d}\sigma) \Big], \end{aligned} \tag{2}$$

where $Q(\boldsymbol{m}) = \|\boldsymbol{m}\|_2^2/n$. The first term of $\mathscr{F}_{\mathrm{TAP}}$ involves the h function, which could be roughly understood as the shifted KL-divergence between a Gaussian distribution and the prior distribution Π. The second term is a quadratic form of $\boldsymbol{m}$ involving the observation matrix $\boldsymbol{Y}$. Since the observation $\boldsymbol{Y}$ is a spiked Gaussian random matrix, the TAP free energy $\mathscr{F}_{\mathrm{TAP}}(\boldsymbol{m}, \Pi)$ for fixed Π and fixed $\boldsymbol{\theta}$ can be regarded as a Gaussian process indexed by $\boldsymbol{m} \in \mathbb{R}^n$.

The conjecture regarding TAP variational inference is that, denoting $\boldsymbol{m}_\star = \arg\min_{\boldsymbol{m}} \mathscr{F}_{\mathrm{TAP}}(\boldsymbol{m}, \Pi)$ for the well-specified prior Π, then $\boldsymbol{m}_\star$ is a consistent estimation of $\boldsymbol{\Theta}_{\mathrm{Bayes}}$, i.e. $\lim_{n\to\infty} \|\boldsymbol{m}_\star \boldsymbol{m}_\star^{\mathsf{T}} - \boldsymbol{\Theta}_{\mathrm{Bayes}}\|_{\mathsf{F}}^2/n^2 \to 0$ almost surely. However, due to the non-convexity of the function $\mathscr{F}_{\mathrm{TAP}}$, it is challenging to verify such a conjecture. Furthermore, it remains an interesting question of analyzing the landscape properties and providing algorithmic guarantees for the TAP-free energy.

Prior work, limitation, and challenges. The TAP free energy was first proposed to approximate the Gibbs average of the spin glass model, a statistical physics model. For spin glasses, the validity of the TAP free energy and their relation to the Gibbs measure have been extensively studied—see for example [81–84] in the physics literature, and [30, 85–91] for rigorous mathematical results. The proposed research studies the TAP free energy for statistical models with a planted signal, which was less known in the literature.

The technical challenge of analyzing the TAP free energy is primarily due to its non-convex landscape. Many recent works established benign landscape properties and algorithmic guarantees for a variety of statistical models [92–95]. Their analyses relied on the uniform convergence of the gradient and Hessian [94] and required the assumption that the signal-to-noise ratio is

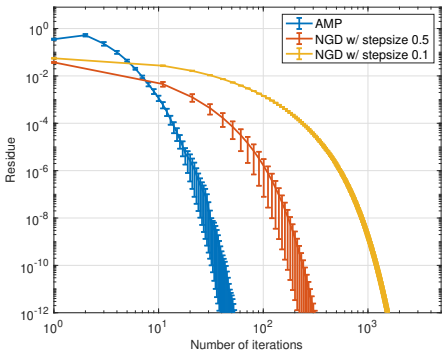

Figure 1. Numerical simulations of optimizing the TAP free energy of $\mathbb{Z}_2$ synchronization. Signal-to-noise ratio is chosen to be $\lambda = 1.5$. The dimension $n = 500$.

strong enough. Unfortunately, such an approach cannot be applied to models in the weak signal regime, including the spiked matrix model. Instead, in the proposed research, we will analyze the landscape of the TAP free energy using techniques for studying Gaussian processes, including the Kac-Rice formula [33, 44] and Gaussian comparison inequalities [36, 37, 39].

Preliminary results. Our preliminary results [45, 76] studied the TAP free energy of the $\mathbb{Z}_2$ synchronization problem [96, 97], which is a special case of the spiked matrix model. In $\mathbb{Z}_2$ synchronization, we specialize the prior $\Pi = \mathrm{Unif}(\{1,-1\})$ to be the Rademacher distribution so that the TAP free energy can be simplified as

$$\begin{aligned} \mathscr{F}_{\mathrm{TAP}}(\boldsymbol{m}) &= -\sum_{i=1}^{n} \mathsf{h}(m_i) - \frac{\lambda}{2}\langle \boldsymbol{m}, \boldsymbol{Y}\boldsymbol{m}\rangle - \frac{n\lambda^2}{4}(1-Q(\boldsymbol{m}))^2, \\ \mathsf{h}(m) &= -\frac{1+m}{2}\log\Big(\frac{1+m}{2}\Big) - \frac{1-m}{2}\log\Big(\frac{1-m}{2}\Big). \end{aligned} \tag{3}$$

We proved benign global and local landscape properties of $\mathscr{F}_{\mathrm{TAP}}$ of the $\mathbb{Z}_2$ model and derived efficient optimization algorithms for finding the global minimum. In the following, we give an itemized list of the results derived in these two preliminary work [45, 76].

- **No spurious local minimizer in an intermediate region.** In [45], we derived an upper bound for the expected number of critical points of $\mathscr{F}_{\mathrm{TAP}}$ in sub-regions of the domain $(-1,1)^n$. Using this result, for $\lambda > \lambda_0$ a large enough absolute constant, it was shown that the global minimizer $\boldsymbol{m}_\star$ of $\mathscr{F}_{\mathrm{TAP}}$ satisfies $\mathbb{E}[\|\boldsymbol{m}_\star\boldsymbol{m}_\star^{\mathsf{T}} - \boldsymbol{\Theta}_{\mathrm{Bayes}}\|_{\mathsf{F}}^2]/n^2 \to 0$, and that this holds more generally for any critical point $\boldsymbol{m}$ of $\mathscr{F}_{\mathrm{TAP}}$ in the domain
$$\mathscr{S} = \{\boldsymbol{m} \in (-1,1)^n : \mathscr{F}_{\mathrm{TAP}}(\boldsymbol{m}) < -\lambda^2/3\}.$$

- **Existence, uniqueness, and convexity of TAP local minimizer.** In [76], we showed that, for any $\lambda > 1$, there exists a local minimizer $\bar{\boldsymbol{m}}_\star$ such that $\|\bar{\boldsymbol{m}}_\star\bar{\boldsymbol{m}}_\star^{\mathsf{T}} - \boldsymbol{\Theta}_{\mathrm{Bayes}}\|_{\mathsf{F}}^2/n^2 \to 0$ in probability. Moreover, $\mathscr{F}_{\mathrm{TAP}}$ is strongly convex in a $\sqrt{\varepsilon n}$-neighborhood of this local minimizer $\bar{\boldsymbol{m}}_\star$, so that $\bar{\boldsymbol{m}}_\star$ is the unique critical point in such a neighborhood. When λ is large, this is identified with the global minimizer $\bar{\boldsymbol{m}}_\star = \boldsymbol{m}_\star$.

- **Convergence guarantees for natural gradient descent and approximate message passing.** In [76], we also considered two natural algorithms for minimizing $\mathscr{F}_{\mathrm{TAP}}$: the approximate message passing (AMP) algorithm [31, 32] which takes the form
$$\boldsymbol{m}^{k+1} = \tanh(\lambda\boldsymbol{Y}\boldsymbol{m}^k - \lambda^2[1-Q(\boldsymbol{m}^k)]\boldsymbol{m}^{k-1}), \tag{AMP}$$
and the natural gradient descent (NGD) algorithm [98] with step size $\eta > 0$, which takes the form
$$\boldsymbol{m}^{k+1} = \tanh(\boldsymbol{h}^{k+1}), \qquad \boldsymbol{h}^{k+1} = (1-\eta)\boldsymbol{h}^k + \eta(\lambda\boldsymbol{Y}\boldsymbol{m}^k - \lambda^2[1-Q(\boldsymbol{m}^k)]\boldsymbol{m}^k). \tag{NGD}$$
For any $\lambda > 1$, we proved that NGD achieves linear convergence to $\boldsymbol{m}_\star$ from any initialization within a $\sqrt{\varepsilon n}$-neighborhood. This initialization may be obtained by first performing a fixed number of iterations of AMP, thus yielding a polynomial-time algorithm for computing $\boldsymbol{m}_\star$. Moreover, for λ large enough, we proved that both AMP and NGD alone exhibit linear convergence to $\boldsymbol{m}_\star$ from a spectral initialization. We emphasize that this convergence of AMP and NGD is established in the sense $\lim_{k\to\infty}\boldsymbol{m}^k = \boldsymbol{m}_\star$ for fixed dimension n, which cannot be shown using the state evolution analyses [32, 42, 99, 100]. We illustrate the convergence of these algorithms on a fixed problem instance using numerical simulations in Figure 1.

We proved these results using the techniques for analyzing Gaussian processes, including the Kac-Rice formula and Gaussian comparison inequalities. We believe these techniques can be generalized to study the geometric and algorithmic properties of non-convex objective functions of other statistical models.

Proposed research. We propose the following tasks for investigation, built upon our preliminary results.

- **Spiked matrix model with a general prior.** In the proposed research, and together with my collaborators Michael Celentano, Zhou Fan, and Andrea Montanari, we will apply similar techniques to study the landscape of the TAP free energy of the spiked matrix model with a general prior Π. We will first focus on the scenario when the prior Π is well-specified. The exact asymptotics of the mean squared error of the Bayes estimator in this model has been derived in a series of earlier works [73, 97, 101–103]. It was shown in these works that for general symmetric prior Π, there are two interesting phase transition thresholds of the signal-to-noise ratio λ. One interesting threshold λ_{IT} is the information theoretical threshold, and when $\lambda < \lambda_{\mathrm{IT}}$, the Bayes estimator performs no better than the zero estimator. The other interesting threshold λ_c is the computational threshold, and when $\lambda > \lambda_c$, the AMP algorithm with spectral initialization will converge to the Bayes estimator in the weak sense.

  We conjecture that, whenever $\lambda > \lambda_{\mathrm{IT}}$, the TAP free energy will have a unique local minimizer $\bar{\boldsymbol{m}}_\star$ (up to $\pm$ sign) near the Bayes estimator $\boldsymbol{\Theta}_{\mathrm{Bayes}}$, which is also the global minimizer and will have a strongly convex neighborhood. Moreover, whenever $\lambda > \lambda_0$ for some large constant λ_0, we can use the NGD algorithm or the AMP algorithm coupled with spectral initialization to find the global minimizer $\boldsymbol{m}_\star$ in the sense that $\lim_{k\to\infty} \boldsymbol{m}^k = \boldsymbol{m}_\star$. It is our aim to establish this conjecture formally.

- **Spiked matrix model with a misspecified prior.** In most scientific applications, the prior knowledge of the statistician will not be fully accurate. When the prior Π' used to solve the Bayesian inference problem is different from the data-generating prior Π, statisticians hope that our inferential procedures will not be completely unreasonable. We are optimistic about this hope, and we believe that, when the KL-divergence between Π and Π' is smaller than some constant ε, the global minimizer of the TAP free energy with prior Π' will still correspond to the Bayes posterior mean under prior Π', which should be close to the Bayes estimator with the well-specified prior Π. Furthermore, there should be similar benign geometric and algorithmic guarantees for the TAP free energy. However, to show these results theoretically, there are quite a few technical challenges.

  One particular challenge is to establish the exact asymptotics of the mean squared error of the Bayes estimator under the misspecified prior Π'. Due to the prior misspecification, the Nishimori identity, as an important equation in proving the exact asymptotics of Bayes estimators as in [73, 101], does not hold anymore. We need to develop more theoretical techniques to address this challenge.

- **Empirical Bayes approach for estimating the prior.** When the prior Π is unknown, it is a natural idea to use the data to estimate the prior in order to improve the estimation accuracy. This idea has been explored in [104] for the spiked matrix model, in which the authors estimate the prior using the non-parametric maximum likelihood estimator applied to the AMP iterates. Such a method for estimating the prior is somewhat ad hoc and cannot be easily generalized. This research thrust proposes an alternative empirical Bayes method to estimate the prior Π: we jointly estimate the Bayes posterior mean $\boldsymbol{m}_\star$ and the Bayes prior $\widehat{\Pi}$ by solving the optimization problem

$$(\boldsymbol{m}_\star, \widehat{\Pi}) = \arg\min_{\boldsymbol{m},\Pi} \mathscr{F}_{\mathrm{TAP}}(\boldsymbol{m}, \Pi). \tag{4}$$

  The method above is inspired by the following intuition: for any fixed Π, by the derivation of the TAP free energy, the TAP minimum $\min_{\boldsymbol{m}} \mathscr{F}_{\mathrm{TAP}}(\boldsymbol{m}, \Pi)$ is approximating the negative likelihood function of Π. If this approximation holds uniformly well for every prior Π, then $\widehat{\Pi}$ is approximately the maximum likelihood estimator of Π. To theoretically verify the validity of the proposed approach, we need to analyze the optimization problem (4) using again the aforementioned techniques.

- **Extension to Bayesian linear models.** Extending the TAP free energy analysis, we turn to study the TAP variational inference for Bayesian linear models [105, 106]. Consider the linear model where we have $\boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ which satisfy a linear relationship $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ for some Gaussian noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \mathbf{I}_n)$. We assume a prior on the parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^\mathsf{T} \in \mathbb{R}^d$ where $(\beta_j)_{j\in[d]} \sim_{iid} \Pi$. We further assume that the covariates $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\mathsf{T}$ have independent Gaussian rows $\boldsymbol{x}_i \sim_{iid} \mathcal{N}(\boldsymbol{0}, (1/n)\mathbf{I}_d)$. We will consider the regime in which the sample size n and the dimension d are proportional to each other.

  The TAP free energy of the Bayesian linear model, derived in [107], takes the form

$$\mathscr{F}_{\mathrm{TAP}}(\boldsymbol{m}, \Pi) = -\sum_{i=1}^{d} \mathsf{h}(m_i; Q(\boldsymbol{m}), \Pi) + \frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{m}\|_2^2 - \mathscr{F}_{\mathrm{Cor}}(\boldsymbol{m}), \tag{5}$$

  where h is the same function as defined in (2) and $\mathscr{F}_{\mathrm{Cor}}$ is the Onsager correction term that is a deterministic function of $\boldsymbol{m}$. The function $\mathscr{F}_{\mathrm{TAP}}$ is no longer a Gaussian process and we cannot directly apply the Kac-Rice formula to analyze its landscape. To resolve this problem, we can use the duality representation of the norm to convert $\mathscr{F}_{\mathrm{TAP}}$ into the supremum of a Gaussian process $\mathscr{F}_{\mathrm{TAP}}(\boldsymbol{m}, \Pi) = \sup_{\boldsymbol{u}\in\mathbb{R}^n} \mathscr{G}_{\mathrm{TAP}}(\boldsymbol{m}, \boldsymbol{u}, \Pi)$, where $\mathscr{G}_{\mathrm{TAP}}$ takes the form

$$\mathscr{G}_{\mathrm{TAP}}(\boldsymbol{m}, \boldsymbol{u}, \Pi) = -\sum_{i=1}^{d} \mathsf{h}(m_i; Q(\boldsymbol{m}), \Pi) + \frac{1}{\sigma^2}\left(\langle \boldsymbol{y} - \boldsymbol{X}\boldsymbol{m}, \boldsymbol{u}\rangle - \frac{1}{2}\|\boldsymbol{u}\|_2^2\right) - \mathscr{F}_{\mathrm{Cor}}(\boldsymbol{m}). \tag{6}$$

  Conditional on $\boldsymbol{\beta}$, $\mathscr{G}_{\mathrm{TAP}}$ is a Gaussian process indexed by $(\boldsymbol{m}, \boldsymbol{u})$, where the randomness comes from $\boldsymbol{X}$ and $\boldsymbol{\varepsilon}$. Then the challenge is to analyze the properties of the minimax optimization problem of the Gaussian process $\mathscr{G}_{\mathrm{TAP}}$, and we need to extend the Kac-Rice analysis to handle this problem.

4 Component B: Optimal Bayesian procedures for finite-sample frequentist FDR control

Suppose that we observe a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ of predictor measurements and a response vector $\boldsymbol{y} \in \mathbb{R}^n$, and assume that

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \mathbf{I}_n), \tag{7}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^\mathsf{T}$ and σ^2 are unknown. In this linear model, we consider the false discovery control problem [1] where the null hypotheses to be tested are $H_{0,j} : \beta_j = 0$ for $j \in \mathscr{H} \equiv \{1, 2, \ldots, d\}$. In general, a test statistics maps the data to a binary variable $T_j : \mathbb{R}^{n\times d} \times \mathbb{R}^d \subseteq \{0, 1\}$, where $T_j = 1$ indicates that we reject the j'th null hypothesis. Define the false positive proportion and the true positive proportion of the test $\boldsymbol{T} = (T_1, \ldots, T_d)^\mathsf{T}$ as

$$\mathrm{FDP}(\boldsymbol{T}) = \frac{\#\{j : T_j = 1, \beta_j = 0\}}{\#\{j : T_j = 1\}}, \qquad \mathrm{TPP}(\boldsymbol{T}) = \frac{\#\{j : T_j = 1, \beta_j \neq 0\}}{\#\{\beta_j \neq 0\}}. \tag{8}$$

A good test is one for which TPP is large and FDP is small, meaning that the test is able to separate non nulls from nulls.

If we assume a prior upon the unknown parameters, the multiple testing problem can be cast into the Bayesian decision-theoretic framework [5, 13, 61]. Given a prior $\boldsymbol{\beta} \sim \boldsymbol{\Pi}$, we can maximize the Bayes true positive rate BTPR while constraining the Bayes false discovery rate BFDR at level α,

$$\mathrm{BTPR}(\boldsymbol{T}; \boldsymbol{\Pi}) = \mathbb{E}_{\boldsymbol{\beta}\sim\boldsymbol{\Pi}}\mathbb{E}_{(\boldsymbol{X},\boldsymbol{y})\sim\mathbb{P}_{\boldsymbol{\beta}}}[\mathrm{TPP}(\boldsymbol{T})], \quad \mathrm{BFDR}(\boldsymbol{T}; \boldsymbol{\Pi}) = \mathbb{E}_{\boldsymbol{\beta}\sim\boldsymbol{\Pi}}\mathbb{E}_{(\boldsymbol{X},\boldsymbol{y})\sim\mathbb{P}_{\boldsymbol{\beta}}}[\mathrm{FDP}(\boldsymbol{T})]. \tag{9}$$

The optimal solution can be easily derived from the Bayesian decision theory [13, 60], which truncates the local FDR [5]

$$T_j(\boldsymbol{X}, \boldsymbol{y}; t) = 1\{\mathbb{P}(\beta_j = 0 | \boldsymbol{X}, \boldsymbol{y}) \leq t\}, \tag{10}$$

at some (data-dependent) threshold t. Unfortunately, the Bayesian framework is often sensitive to model misspecification and prior mismatch despite being tractable and straightforward.

On the other hand, without depending on the prior, frequentist methods aim to control the frequentist false discovery rate

$$\mathrm{FDR}(\boldsymbol{T};\mathbb{P}_{\boldsymbol{\beta}}) = \mathbb{E}_{(\boldsymbol{X},\boldsymbol{y})\sim\mathbb{P}_{\boldsymbol{\beta}}}[\mathrm{FDP}(\boldsymbol{T})] \tag{11}$$

under an unknown distribution $\mathbb{P}_{\boldsymbol{\beta}}$. Furthermore, a few recent procedures, including the knockoff procedure [11, 12] and the conditional randomization test (CRT) [12, 66], control the frequentist FDR under the model-X setting: the data comes from a joint distribution $(\boldsymbol{x}_i, y_i) \sim_{iid} \mathscr{L}((\boldsymbol{x}, y))$, whereas the test statistics can access the marginal distribution $\mathscr{L}(\boldsymbol{x})$. These procedures are finite-sample valid regardless of the conditional model $\mathscr{L}(y|\boldsymbol{x})$ so that they are robust with respect to misspecified conditional models $\mathscr{L}(y|\boldsymbol{x})$. However, it remains an open question of how to optimally incorporate the prior information into these procedures to improve the power. This is the aforementioned "dual objective" challenge in the FDR control task.

Prior work, limitation, and challenges. A few recent works [67–69] carried out the power analysis of model-X valid procedures including the knockoff procedure and the CRT. For example, [68] derived the exact asymptotics of the FDP and TPP of the knockoff procedure applied to LASSO coefficient difference statistics. However, the procedures considered in this work are not near-optimal despite controlling the frequentist FDR. Indeed, in the presence of a well-specified prior, the thresholded LASSO procedure will be sub-optimal compared to the procedure of truncating the local FDR (10). Furthermore, the knockoff procedure will lose power compared to the oracle thresholded LASSO procedure [68].

Combining the model-X valid procedures with the local FDR statistics is a natural idea to resolve the "dual objective" challenge. However, it is technically difficult to analyze the power of such procedures. We cannot apply the approximate message passing approach, which is the theoretical tool to derive the exact asymptotics of the thresholded LASSO statistics [67, 68]. We will instead resolve this technical difficulty using interpolation methods for analyzing the exact asymptotics of Bayesian models [105, 106].

Preliminary work. We consider the test statistics $\boldsymbol{T}_{\mathsf{TPoP}} = (T_{\mathsf{TPoP},j})_{j\in[d]}$, where each coordinate truncates the local FDR with threshold t

$$T_{\mathsf{TPoP},j}(\boldsymbol{X},\boldsymbol{y}) = 1\{\mathbb{P}(\beta_j = 0|\boldsymbol{X},\boldsymbol{y}) \le t\}. \tag{12}$$

The threshold t will be calibrated using the BFDR nominal level α and will be specified later. We give the procedure $\boldsymbol{T}_{\mathsf{TPoP}}$ an acronym TPoP standing for truncating the posterior probability. Prior works [13, 60] have shown that such a test statistics has the largest BTPR given BFDR controlled at a specific level.

An ongoing work of the PI (with two Berkeley Ph.D. students, Taejoo Ahn and Licong Lin) derives the exact asymptotics of the FDP and TPP of the TPoP procedure under the following assumptions of the Bayesian linear model. These assumptions are natural for power analysis of FDR control procedures and are mostly the same as in [67, 68].

Assumption 1. *Consider the linear model (7). We assume that $\boldsymbol{\beta} = (\beta_1,\dots,\beta_d)^{\mathsf{T}}$ has independent prior on the coordinates $\beta_j \sim_{iid} \Pi$, where $\Pi = \pi_0\delta_0 + (1-\pi_0)\Pi_\star$ for some sparsity parameter $0 < \pi_0 < 1$ and some general distribution $\Pi_\star \in \mathscr{P}(\mathbb{R})$. We further assume that $\boldsymbol{X} = (\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)^{\mathsf{T}}$ has independent rows with isotropic Gaussian distribution $\boldsymbol{x}_i \sim_{iid} \mathcal{N}(\boldsymbol{0},(1/n)\mathbf{I}_d)$.*

The following conjecture is derived using the replica method, which gives the asymptotic FDP and TPP of the TPoP procedure of the Bayesian linear model as in Assumption 1.

Conjecture 1. *Consider the TPoP procedure $\boldsymbol{T}_{\mathsf{TPoP}} = (T_{\mathsf{TPoP},j})_{j\in[d]}$ as defined in (12). Define*

$$\mathscr{E}(h;\tau) = \mathbb{E}_{(\beta,G)\sim\Pi\times\mathcal{N}(0,1)}[\beta|\beta+\tau G = h], \qquad \mathscr{P}(h;\tau) = \mathbb{P}_{(\beta,G)\sim\Pi\times\mathcal{N}(0,1)}[\beta = 0|\beta+\tau G = h]. \tag{13}$$

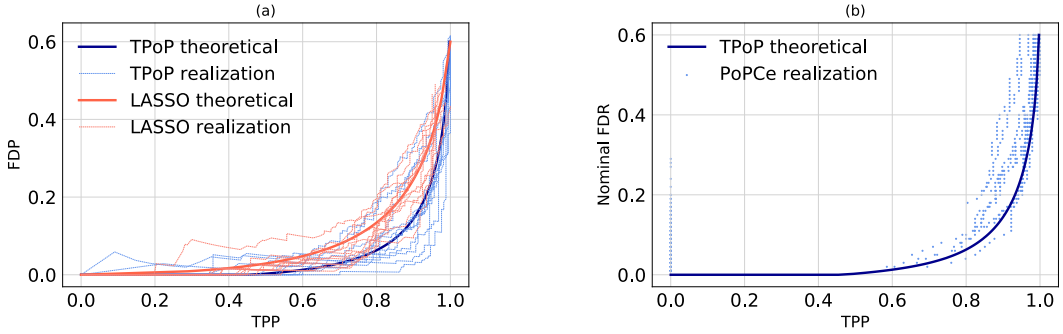

Figure 2. FDP and TPP tradeoff curves under the Bayesian linear model with the three delta prior $\Pi = 0.2\delta_1 + 0.2\delta_{-1} + 0.6\delta_0$. We took $d = 400$ and $n = 500$. Left: Theoretical predictions and numerical simulations of the TPP and FDP of the TPoP and the thresholded LASSO procedure. Right: Theoretical predictions and numerical simulations of the TPP and the nominal FDR level of the PoPCe procedure. The TPoP and PoPCe are approximately computed using the TAP variational inference method.

For some $\delta \in (0,\infty)$, let $\tau_\star$ be the largest non-negative solution of

$$\tau^2 = \sigma^2 + \frac{1}{\delta}\mathbb{E}_{(\beta,Z)\sim\Pi\times\mathcal{N}(0,1)}[(\beta - \mathscr{E}(\beta+\tau Z;\tau))^2]. \tag{14}$$

Then in the limit of $n,d\to\infty$, and $n/d\to\delta$, we have almost surely

$$\lim_{n,d\to\infty} \mathrm{FDP}(\boldsymbol{T}_{\mathsf{TPoP}}) = \mathbb{P}_{(\beta,Z)\sim\Pi\times\mathcal{N}(0,1)}[\beta = 0 | \mathscr{P}(\beta+\tau_\star Z;\tau_\star) \geq t], \tag{15}$$

$$\lim_{n,d\to\infty} \mathrm{TPP}(\boldsymbol{T}_{\mathsf{TPoP}}) = \mathbb{P}_{(\beta,Z)\sim\Pi\times\mathcal{N}(0,1)}[\mathscr{P}(\beta+\tau_\star Z;\tau_\star) \geq t | \beta \neq 0]. \tag{16}$$

Recall that there is an undetermined threshold t in the TPoP procedure. If we aim to control the BFDR at level α, we can determine $t = t_\star(\alpha)$ where $t_\star(\alpha)$ solves the equation by taking the right-hand side of (15) equal to α

$$\mathbb{P}_{(\beta,G)\sim\Pi\times\mathcal{N}(0,1)}(\beta = 0 | \mathscr{P}(\beta+\tau_\star G) \leq t) = \alpha. \tag{17}$$

Then, such a procedure will asymptotically control the BFDR at level α, implied by Conjecture 1. To compare the power of TPoP and the thresholded LASSO procedure, we plot their TPP-FDP tradeoff curves in Figure 2 (a). The thick curves are the theoretical predictions given by Conjecture 1 and by theorems in [68]. The thin curves are the numerical simulations with finite dimension d. The figure shows that the TPoP procedure is much more powerful than the thresholded LASSO procedure.

However, despite controlling the BFDR, the TPoP procedure does not control the frequentist FDR under general data distribution. Suppose that the data satisfies $(\boldsymbol{x}_i, y_i) \sim_{iid} \mathscr{L}((\boldsymbol{x},y))$, where $\mathscr{L}((\boldsymbol{x},y))$ is a general data distribution which may not be the Bayesian linear model. In this misspecified model, we hope to test the hypotheses that $y \perp x_j | \boldsymbol{x}_{-j}$ for each $j \in [d]$ and control the frequentist FDR. However, it is impossible to achieve such a goal without any knowledge of $\mathscr{L}((\boldsymbol{x},y))$ [108]. One way to make the goal feasible is to consider the model-X setting, similar to [12, 66], in which the covariate distribution $\mathscr{L}(\boldsymbol{x})$ is known, and the conditional distribution $\mathscr{L}(y|\boldsymbol{x})$ can be arbitrary.

Our preliminary result proposes Algorithm 1, which controls the frequentist FDR under arbitrary models $\mathscr{L}((\boldsymbol{x},y))$ and is as powerful as the TPoP procedure under the Bayesian linear model satisfying Assumption 1. This procedure is a combination of the TPoP procedure [9], the conditional randomization test (CRT) [12], and the eBH procedure [65]. We give it an acronym PoPCe (Posterior Probability, CRT, eBH).

Our next theorem shows that PoPCe controls the frequentist FDR under arbitrary data distribution.

Algorithm 1 The PoPCe procedure $\boldsymbol{T}_{\mathsf{PoPCe}}$

Require: Dataset $\{(\boldsymbol{x}_i, y_i)\}_{i\in[n]}$; FDR control level $\alpha \in (0,1)$; distribution of covariates $\mathscr{L}(\boldsymbol{x})$; null proportion estimate π_0; prior estimate Π and noise estimate σ^2; hyperparameter $K \in \mathbb{N}$.

1: (Posterier Probability) Compute the local FDR $P_j = \mathscr{P}_j(\boldsymbol{X}, \boldsymbol{y})$, where $\mathscr{P}_j(\boldsymbol{X}, \boldsymbol{y}) = \mathbb{P}(\beta_j = 0 | \boldsymbol{X}, \boldsymbol{y})$.

2: (Conditional Randomization Test) Generate the p-values using the conditional randomization test, i.e., sample $\tilde{\mathbf{x}}_j^{(k)} = (\tilde{x}_{1j}^{(k)}, \ldots, \tilde{x}_{nj}^{(k)})^{\mathsf{T}} \in \mathbb{R}^n$ where $\tilde{x}_{ij}^{(k)} \sim \mathscr{L}(x_j | \boldsymbol{x}_{-j} = \boldsymbol{x}_{i,-j})$ independently. Then compute the p-values $p_j = (1 + \sum_{k=1}^{K} 1\{P_j \geq \mathscr{P}_j(\boldsymbol{X}_{-j}, \tilde{\mathbf{x}}_j^{(k)}, \boldsymbol{y})\})/(1+K)$.

3: (eBH procedure) Convert the p-values into e-values by $e_j = 1_{p_j < q}/q$ for some pre-determined threshold q. Reject the hypotheses corresponding to the $\hat{k}$ largest e-values, where $\hat{k} = \max\{k : \frac{\pi_0 d}{k e_{(k)}} \leq \alpha\}$.

Theorem 1 (Frequentist FDR control of PoPCe)**.** *Suppose that $(\boldsymbol{x}_i, y_i) \sim_{iid} \mathscr{L}((\boldsymbol{x}, y))$ for arbitrary joint distribution $\mathscr{L}((\boldsymbol{x}, y))$ with $\pi_0 d$ number of null-hypothesis. Let (Π, σ^2) be arbitrary prior estimate and noise estimate that are independent of the dataset $\{(\boldsymbol{x}_i, y_i)\}_{i\in[n]}$. Then* PoPCe *controls the FDR (11)*

$$\mathrm{FDR}(\boldsymbol{T}_{\mathsf{PoPCe}}, \mathscr{L}((\boldsymbol{x}, y))) \leq \alpha.$$

The conjecture below shows that, for a specific choice of the parameter q, the PoPCe procedure nearly attains the same power as the TPoP procedure, so that it is near-optimal when the model is well-specified. This conjecture is again derived using the heuristic replica method, and we aim to prove it in this thrust.

Conjecture 2. *Suppose that $\{(\boldsymbol{x}_i, y_i)\}_{i\in[n]}$ satisfies Assumption 1 and that we use the true prior and noise level (Π, σ^2). We take $q = \Psi(t_\star(\alpha)) - \varepsilon$ for some small $\varepsilon > 0$, where Ψ is the cumulative distribution function of $\mathscr{P}(\tau_\star G)$ when $G \sim \mathbb{N}(0,1)$, and $t_\star(\alpha)$ solves (17). We take the hyperparameter $K = \infty$. Then the TPP of the* PoPCe *procedure is asymptotically the same as* TPoP *in the limit of $n, d \to \infty$, and $n/d \to \delta$*

$$\lim_{\varepsilon \to 0} \lim_{n,d\to\infty} \mathrm{TPP}(\boldsymbol{T}_{\mathsf{PoPCe}}) = \lim_{n,d\to\infty} \mathrm{TPP}(\boldsymbol{T}_{\mathsf{TPoP}}).$$

Recall that TPoP is optimal in the Bayes sense [13, 60]: it maximizes the BTPR among all the procedures with BFDR control at level α. Since Conjecture 2 shows that PoPCe has the same power as TPoP under the Bayes linear model, PoPCe also nearly maximizes the BTPR among all the procedures with FDR control at level α. In Figure 2 (b), we numerically verified that PoPCe attains the same power as TPoP.

Proposed research. The first task of this research thrust is to prove Conjecture 1 and 2. The PI plans to work on this task together with two Berkeley Ph.D. students Taejoo Ahn and Licong Lin. We have verified these conjectures using extensive numerical simulations. To prove these conjectures, we will first try the leave-one-out approach and interpolation methods [30, 74], which were shown to be useful in proving the exact asymptotics of the minimum mean-squared error of Bayesian linear models [73–75].

Beyond proving these conjectures, we propose below further tasks to be investigated in this thrust.

- **Empirical Bayes approach for estimating the prior.** In modern applications, we usually do not precisely know the prior but instead, estimate the prior using the data. In this task, we propose to investigate the optimal approach for estimating the prior in the PoPCe procedure. If we naively use the same dataset to estimate the prior, the PoPCe procedure will not be frequentist valid: p_j's will not be valid p-values since we are reusing the samples. If we naively use the sample splitting method, we may lose power since we only use partial samples for estimating the prior and partial samples for the PoPCe procedure.

  A more sophisticated idea uses the leave-one-out approach. However, this leave-one-out approach is not leaving one sample out but leaving one covariate out. More specifically, when we compute the j'th p-value p_j, we plug in the prior $\widehat{\Pi}_{-j}$ estimated from the dataset $\{(y_i, \boldsymbol{x}_{i,-j})\}_{i\in[n]}$, where $\boldsymbol{x}_{i,-j} \in \mathbb{R}^{d-1}$ is the

i'th sample leaving out the j'th coordinate. As a consequence, p_j conditional on $\{(y_i, \boldsymbol{x}_{i,-j})\}_{i\in[n]}$ will be a valid p-value, so that we can guarantee the frequentist validity of the procedure. Furthermore, since the leave-one-out approach uses almost all the samples for estimating the prior and for the PoPCe procedure, we will not lose much power. The proposed research is to justify this idea using theories and numerical simulations. Finally, we remark that the leave-one-out approach will be computationally heavy, and it is an interesting question to make it computationally efficient.

- **Near optimal procedure for Bayesian linear model with anisotropic design.** The isotropic design assumption as in Assumption 1 is very strong. A relatively weaker assumption is the anisotropic design, i.e., $\boldsymbol{x}_i \sim_{iid} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ for some known covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$. The proposed research aims to develop near-optimal FDR controlling procedures under the anisotropic design Bayesian linear model.

  Recall that Theorem 1 guarantees that PoPCe controls the frequentist FDR under arbitrary distribution $\mathscr{L}((\boldsymbol{x}, y))$, which includes the anisotropic design Bayesian linear model. However, PoPCe does not give the optimal BTPR under the anisotropic design assumption. To establish an anisotropic design analog of Conjecture 2, we need first to calculate the asymptotic limit of the TPP and FDP of the TPoP procedure in the anisotropic setting. We propose to use the replica method to heuristically derive such an asymptotic limit. This approach was adopted in [10] to heuristically derive the exact asymptotic of the LASSO procedure under anisotropic design, which was later proved in [109]. After deriving the asymptotic limit of the TPoP procedure, a further challenge is to combine the TPoP with the conditional randomization test or its variants to make it frequentist valid under arbitrary distribution.

- **Near optimal frequentist FDR control in Bayesian generalized linear model.** We propose to extend the above analysis to Bayesian generalized linear models. Consider the generalized linear model where we observe for $i = 1, 2, \ldots, n$ the measurements $y_i \sim \pi(y|\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta})$ with prior $(\beta_j)_{j\in[d]} \sim_{iid} \Pi$. Here $\pi(y|\theta)$ is a probability distribution over y indexed by θ, for example, the Bernoulli distribution. We further assume that the covariates are independent Gaussian random vectors $\boldsymbol{x}_i \sim_{iid} \mathcal{N}(\boldsymbol{0}, (1/n)\mathbf{I}_d)$. In this model, the TPoP procedure is still the optimal BFDR control procedure, and we will first derive its asymptotic FDP and TPP. Next, we will consider to design a near-optimal frequentist FDR control procedure by adapting the PoPCe procedure to this model.

  Finally, we propose to investigate the power of the Knockoff procedure [11, 12] and mirror statistics procedures [110, 111] combined with the local FDR statistics. These are two alternative methods for FDR control. It is an interesting question whether these procedures or their variants can achieve near-optimal power in Bayesian generalized linear models.

5 Component C: Bayesian methods and distribution-free predictive inference.

We consider the task of constructing prediction intervals in regression problems. Let $\{(\boldsymbol{x}_i, y_i)\}_{i\in[m]} \sim_{iid} \mathscr{L}((\boldsymbol{x}, y))$ be independent and identically distributed as $(\boldsymbol{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ for some arbitrary data distribution $\mathscr{L}((\boldsymbol{x}, y))$. The task is to construct a prediction interval $C(\boldsymbol{x})$ of y with a valid coverage guarantee. At the same time, we hope that the prediction interval is informative and as small as possible.

Conformal prediction methods [2] treat base predictive inference procedures as black boxes, wrap around these procedures, and provide a frequentist marginal coverage guarantee upon the resulting procedures under only general exchangeability assumptions. To illustrate the idea of conformal prediction, we consider the splitting conformal prediction method as an example. Formally, let $\{\mathscr{I}_{\text{train}}, \mathscr{I}_{\text{cal}}\}$ form a partition of $\{1, \cdots, m\}$, and let $n = |\mathscr{I}_{\text{cal}}|$. Without loss of generality, we take $\mathscr{I}_{\text{cal}} = \{1, \ldots, n\}$. We allow the statistician to fit a sequence of nested prediction intervals $\{C_t\}_{t\in\mathscr{T}}$ on the training set $\mathscr{I}_{\text{train}}$ using an arbitrary procedure. For example, the statistician can take $C_t(\boldsymbol{x}) = [f_l(\boldsymbol{x}) - t, f_u(\boldsymbol{x}) + t]$ for some trained function

f_l and f_u [17]. Then, roughly speaking, splitting conformal prediction is choosing the parameter $\hat{t}$ according to a constrained optimization problem using the calibration dataset

$$\hat{t} = \arg\min_{t \geq 0} \frac{1}{n} \sum_{i \in \mathscr{I}_{\text{cal}}} |C_t(\boldsymbol{x}_i)|, \quad \text{subject to} \quad \frac{1}{n} \sum_{i \in \mathscr{I}_{\text{cal}}} 1\{y_i \in C_t(\boldsymbol{x}_i)\} \geq 1 - \alpha. \tag{18}$$

We remark that the original splitting conformal prediction [2] was not formulated as (18) but sometimes is the same as the latter. Indeed, we present the formulation (18) to highlight that conformal prediction is a method that explicitly calibrates the threshold. Theoretically, under iid assumptions with arbitrary distribution $\mathscr{L}((\boldsymbol{x}, y))$, it can be guaranteed that $\mathbb{P}(y \in C_{\hat{t}}(\boldsymbol{x})) \geq 1 - \alpha$, where the probability is over the randomness in $\{(\boldsymbol{x}_i, y_i)\}_{i \in [m]}, (\boldsymbol{x}, y) \sim \mathscr{L}((\boldsymbol{x}, y))$.

There is usually an additional property upon conformal prediction methods: on a particular problem instance, if the base prediction interval happens to cover the true response approximately at the nominal level $1 - \alpha$, then the conformalized prediction interval will be almost the same as the base prediction interval. As a consequence, if we hope to construct a prediction interval that is the shortest under a specific distribution $\mathbb{P}_{\boldsymbol{\beta}}$, and has the frequentist coverage guarantee under arbitrary distribution $\mathscr{L}((\boldsymbol{x}, y))$, then we can conformalize a base procedure that satisfies the following condition: under the distribution $\mathbb{P}_{\boldsymbol{\beta}}$, the base prediction interval is the shortest and covers y approximately at the nominal level. However, it remains open questions of how to construct the base procedure that satisfies this condition in specific statistical models $\mathbb{P}_{\boldsymbol{\beta}}$, and how to efficiently apply conformal inference methods to the base procedure without losing power.

To approach this challenge, let us start by taking $\mathbb{P}_{\boldsymbol{\beta}}$ to be the Bayesian linear model with prior $\boldsymbol{\beta} \sim \boldsymbol{\Pi}$, satisfying Assumption 1. We remark that we will rely on the Bayesian linear model to calculate the length of the prediction interval; however, when we talk about valid coverage, data could follow arbitrary distributions.

Proposed research. We propose below tasks to be investigated in this research program.

- **Conformalizing the Bayesian credible prediction interval.** A natural candidate of the minimal length distribution-free procedure is to conformalize the Bayesian credible prediction interval. Bayesian credible prediction interval minimizes the posterior prediction interval length subject to the posterior coverage $1 - \alpha$. More specifically, it solves

$$C_{\text{Bayes}}(\cdot) = \arg\min_{C(\cdot)} \mathbb{E}_{\mathscr{D}_{\text{train}}, (\boldsymbol{x}, y), \boldsymbol{\beta}}[|C(\boldsymbol{x})|], \quad \text{subject to} \quad \mathbb{P}_{\mathscr{D}_{\text{train}}, (\boldsymbol{x}, y), \boldsymbol{\beta}}(y \in C(\boldsymbol{x})) \geq 1 - \alpha,$$

  where the expectation and probability is with respect to the joint distribution of the training data set, the test point, and the Bayes prior. Bayesian decision theory suggests that the solution is the Bayesian credible prediction interval, which gives $C_{\text{Bayes}}(\boldsymbol{x}) = \{y : p(y | \{(\boldsymbol{x}_i, y_i)\}_{i \in \mathscr{I}_{\text{Train}}}, \boldsymbol{x}) \geq t_{\text{Bayes}}\}$, where t_{Bayes} is such that $\mathbb{P}_{\mathscr{D}_{\text{train}}, (\boldsymbol{x}, y), \boldsymbol{\beta}}(y \in C_{\text{Bayes}}(\boldsymbol{x})) = 1 - \alpha$. There is a natural nested sequence of prediction sets, which is $C_t(\boldsymbol{x}) = \{y : p(y | \{(\boldsymbol{x}_i, y_i)\}_{i \in \mathscr{I}_{\text{Train}}}, \boldsymbol{x}) \geq t\}$. Given the nested prediction sets, one can use splitting conformal prediction (18) to calibrate the parameter $\hat{t}$.

  To show that the conformalized Bayesian credible prediction interval has near minimal length, we need to check whether the Bayesian credible prediction interval achieves coverage level $1 - \alpha$ in the asymptotic limit $d, n \to \infty$ and $n/d \to \delta$. We propose to derive the exact asymptotics of $\mathbb{P}_{(\boldsymbol{x}, y)}(y \in C_{\hat{t}}(\boldsymbol{x}) | \mathscr{D}_{\text{Train}}, \boldsymbol{\beta})$. Suppose such a quantity concentrates at level $1 - \alpha$, then the conformalized prediction interval should also have nearly the same length, which is near-optimal.

- **Conformalizing Bayes neural networks.** Bayesian deep learning methods construct prediction sets by assuming that the data is generated by a neural network with a prior on the model parameters [112–120]. It is hard to believe that the Bayesian neural networks can well-specify the data-generating mechanisms,

and such methods do not have frequentist coverage guarantees; however, these methods perform very well empirically on a variety of datasets. Therefore, it is interesting to study the optimal way to combine them with conformal prediction methods to obtain small prediction sets with frequentist coverage guarantees.

We again consider the Bayesian linear model as the data-generating model when we measure the prediction interval length. We further use the Bayesian random features model, as a simple Bayesian neural network with only one single hidden layer, to construct prediction intervals. More specifically, we assume the data is generated from the following Bayesian model

$$y_i = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} a_j \sigma(\langle \boldsymbol{x}_i, \boldsymbol{w}_j \rangle) + \varepsilon_i, \ \boldsymbol{w}_j \sim_{iid} \mathcal{N}(\boldsymbol{0}, (\tau_w^2/d)\mathbf{I}_d), \ \ a_j \sim_{iid} \mathcal{N}(0, \tau_a^2), \ \ \varepsilon_i \sim_{iid} \mathcal{N}(0, \tau_\varepsilon^2), \quad (19)$$

where $\tau_w^2, \tau_a^2, \tau_\varepsilon^2$ are hyper-parameters. To construct a sequence of prediction intervals, we first perform Bayesian inference to derive the posterior distribution $p(y|\{(\boldsymbol{x}_i, y_i)\}_{i\in\mathscr{I}_{\text{Train}}}, \boldsymbol{x})$ by marginalizing over the variables $(a_j, \boldsymbol{w}_j)_{j\in[N]}$ and $(\varepsilon_i)_{i\in[n]}$, and set

$$C_{\tau_w^2,\tau_a^2,\tau_\varepsilon^2,t}(\boldsymbol{x}) = \{y : p(y|\{(\boldsymbol{x}_i, y_i)\}_{i\in\mathscr{I}_{\text{Train}}}, \boldsymbol{x}) \geq t\}. \quad (20)$$

We then apply conformal prediction methods to calibrate the parameters $(\tau_w^2, \tau_a^2, \tau_\varepsilon^2, t)$. Note that here $\{C_{\tau_w^2,\tau_a^2,\tau_\varepsilon^2,t}\}_{\tau_w^2,\tau_a^2,\tau_\varepsilon^2,t}$ is not a nested set, and we cannot apply standard conformal prediction methods. However, we can use the recent technique of [121] to avoid the requirement of a nested set. To analyze the power of conformalized Bayesian random features prediction interval, we propose to calculate $\mathbb{E}_{\boldsymbol{x}}[|C_{\tau_w^2,\tau_a^2,\tau_\varepsilon^2,t}(\boldsymbol{x})||\mathscr{D}_{\text{Train}}, \boldsymbol{\beta}]$ in the asymptotic limit of $n, d, N \to \infty$ and $N/d \to \psi_1 \in (0, \infty)$, $n/d \to \psi_2 \in (0, \infty)$. These calculations can be carried out by using the recent techniques developed by the PI and other researchers [122–125] studying the exact asymptotics of random features models.

- **Comparing to quantile regression with random features models.** Since Bayesian neural networks will always misspecify the data generating mechanism, why is it a good method for predictive inference? Can we use quantile regression with neural networks to achieve the same performance as Bayesian neural networks? In this task, we choose the base prediction interval as the interval generated by regularized quantile regression with random features models

$$\hat{\boldsymbol{a}}_\alpha = \arg\min_{\boldsymbol{a}} \frac{1}{n} \sum_{i=1}^{n} \ell_\alpha(y_i, f(\boldsymbol{x}; \boldsymbol{a})) + \lambda \|\boldsymbol{a}\|_2^2, \qquad f(\boldsymbol{x}; \boldsymbol{a}) = \sum_{j=1}^{N} a_j \sigma(\langle \boldsymbol{x}_i, \boldsymbol{w}_j \rangle) \quad (21)$$

where $\ell_\alpha(t) = -(1-\alpha)t \cdot 1_{t<0} + \alpha t \cdot 1_{t>0}$ is the pinball loss, the standard loss function for quantile regression. The quantile regression prediction interval is then given as $C(\boldsymbol{x}) = [f(\boldsymbol{x}; \hat{\boldsymbol{a}}_{\alpha/2}), f(\boldsymbol{x}; \hat{\boldsymbol{a}}_{1-\alpha/2})]$. We propose to calculate the prediction interval length of the conformalized version of quantile regression in the limit of $n, d, N \to \infty$ and $N/d \to \psi_1$, $n/d \to \psi_2$, and compare it to the prediction interval length of conformalized Bayesian random features models. Our goal is to understand the conditions under which quantile regression performs better and the conditions under which Bayesian methods are preferred.

6 Broader Impacts

The theory and procedures developed in this project will have a direct impact on various science and engineering applications. Variational inference methods are broadly applied in computational biology, neuroscience, computer vision, and natural language processing. Multiple testing problems naturally arise in biological and medical applications. Predictive inference methods are important in risk-sensitive domains such as healthcare and autonomous vehicles. Furthermore, the theoretical tools will be of great interest to other subjects, including probability theory and theoretical computer science.

Course Development. The mean field asymptotics theory, which flourished in the past decade, is becoming a powerful theoretical toolbox in high dimensional statistics. However, few research universities offered statistics courses on this topic. To fill in such a gap, in Spring 2021 at UC Berkeley, the PI developed an advanced-level course, "STAT260 Mean Field Asymptotics in Statistical Learning", which covers statistical physics methods and rigorous mathematical approaches for deriving the exact asymptotics of high dimensional statistical models. The PI has made the lecture notes publicly available online. *Many students and researchers working in related fields have read the lecture notes and told the PI that this is an excellent and well-developed course, from which they learned this toolbox and started their research. A significant proportion of the benefited researchers are from other research institutes and different countries.* The PI will teach a related course in Spring 2022, "STAT210B Theoretical Statistics", which covers non-asymptotic theories in high dimensional statistics. The PI will continue to develop both courses in the future, incorporating the proposed research topics and results into course materials.

Student Mentoring. The PI plans to involve both graduate and undergraduate students to work on the tasks proposed in this research program. The preliminary work of Component B was joint work with two Ph.D. students of the statistics department at UC Berkeley, Taejoo Ahn and Licong Lin, and the PI plans to use a significant part of this grant to support these two students. Taejoo Ahn is in his fourth year and will be able to turn this into an important part of his thesis. Furthermore, part of the proposed research in Component B is under investigation by a female undergraduate student, Mengqi Lin. The PI also plans to involve a Ph.D. student, Nikhil Ghosh, in Component C. Besides working on the research projects, the students will be exposed to a variety of disciplines and develop coding skills useful for data analytics.

Dissemination. The PI will disseminate the proposed research to a variety of communities, including statistics, applied mathematics, machine learning, and computer science. Starting from July 2020, the PI is co-organizing the "One World Seminar Series on the Mathematics of Machine Learning", a weekly online seminar series as a response to the COVID pandemic. This seminar provides an online platform for discussing research advances in mathematics of data science and has attracted attendees from many disciplines and different countries. In Fall 2021, in the Simons Institute for the Theory of Computing at UC Berkeley, the PI is co-organizing a reading group on the topics of "statistical physics methods in statistics", which attracted many experts on the more focused topic. In the future, the PI plans to co-organize invited sessions at various venues such as JSM, SIAM, INFORMS conferences. The PI will actively help organize workshops and summer schools on the related topics at the *Simons Institute* and the *Mathematical Sciences Research Institute*, which are both located at Berkeley.

Outreach to Underrepresented Minorities and Local Communities. Joining UC Berkeley in Fall 2020, the PI has been supervising four undergraduate research projects. One is a female student, and two of them are exchange students. In the two seminars organized by the PI, the statistics seminar at Berkeley statistics department and the "One World Seminar Series on the Mathematics of Machine Learning", the PI managed to invite a significant proportion of speakers from underrepresented groups. The PI also serves as a volunteer of the SMASH (Summer Math and Science Honors) Academy at Berkeley, whose goal is to eliminate the barriers entering STEM faced by underrepresented students. In the future, the PI plans to host international undergraduates (in particular, underrepresented groups) as summer research interns through the internship programs at Berkeley, and to actively participate in outreach events like the *Cal Day*, which attract thousands of middle school students and introduce them to basic data and statistics.

7 Results from prior NSF support: Not applicable